Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems

Subbarao Kambhampati, Sarath Sreedharan, Mudit Verma, Yantian Zha, Lin Guan
School of Computing & AI, Arizona State University

Abstract

Despite the surprising power of many modern AI systems that often learn their own representations, there is significant discontent about their inscrutability and the attendant problems in their ability to interact with humans. While alternatives such as *neuro-symbolic* approaches have been proposed, there is a lack of consensus on what they are about. There are often two independent motivations (i) symbols as a *lingua franca* for human-AI interaction and (ii) symbols as (system-produced) abstractions use in its internal reasoning. The jury is still out on whether AI systems will need to use symbols in their internal reasoning to achieve general intelligence capabilities. Whatever the answer there is, the need for (human-understandable) symbols in human-AI interaction seems quite compelling. Symbols, like emotions, may well not be *sine qua non* for intelligence *per se*, but they will be crucial for AI systems to interact with us humans–as we can neither turn off our emotions nor get by without our symbols. In particular, in many human-designed domains, humans would be interested in providing explicit (symbolic) knowledge and advice–and expect machine explanations in kind. This alone requires AI systems to at least do their I/O in symbolic terms. In this *blue sky* paper, we argue this point of view, and discuss research directions that need to be pursued to allow for this type of human-AI interaction.

1 Introduction

AI research community is grappling with an ongoing tussle between symbolic and non-symbolic approaches–with the former using representations (and to some extent, knowledge) designed by the users, but are often outperformed by the latter that *learn their own representations*, but at the expense of inscrutability to humans in the loop. While *neuro-symbolic systems* have received attention in some quarters (Garcez et al. 2019; De Raedt et al. 2019), the jury is still out on whether or not AI systems need internal symbolic reasoning to reach human-level intelligence. There are however compelling reasons for AI systems to communicate (take advice or provide explanations) from humans in essentially symbolic terms. After all, the alternatives would be either for the humans to understand the internal (learned) representations of the AI systems–which seems like a rather poor way for us to design *our* future; or for both humans and AI systems to essentially depend on the lowest common substrate they can exchange raw data–be they images, videos or general *space time signal tubes* (heretofore referred to as *STST*).

While STSTs–in particular saliency regions over images–have been used in the machine learning community as a means to either advice or interpret the operation of AI systems (Greydanus et al. 2018; Zhang et al. 2020), we contend that they will not scale to human-AI interaction in more complex sequential decision settings involving both tacit and explicit task knowledge (Kambhampati 2021). This is because exchanging information via STSTs presents high cognitive load for humans–which is what perhaps lead humans to evolve a symbolic language in the first place.[1]

In this paper, we argue that orthogonal to the issue of whether AI systems use internal symbolic representations, AI systems need to develop local symbolic representations that are interpretable to humans in the loop, and use them to take advice and/or give explanations for their decisions. The underlying motivations here are that human-AI interaction should be structured *for the benefit of the humans*–thus the communication should be in terms that make most sense to humans. This argues for the inclusion of a symbolic interface,[2] especially in terms of symbols that already have meaning to the humans in the loop (that is, these cannot just be internal symbolic abstractions that the machine may have developed for its own computational efficiency. Our argument is not that human-AI interaction must be exclusively in symbolic means–but that it is crucial to also support a symbolic interface. As argued in (Kambhampati 2021), AI systems' inability to take explicit knowledge-based advice, or provide interpretable explanations are at the root of many of the ills of the modern AI systems that learn their own internal representations.

Supporting such a *lingua franca* symbolic interface brings

[1] The urge to use symbolic representations for information exchange seems so strong that humans even develop symbolic terms for speaking about even purely tacit tasks (e.g. *pitch and roll* in basket ball).

[2] The oft-repeated "System 1/System 2" architectural separation, on the other hand, doesn't strictly necessitate or lead to symbols that will serve as *lingua franca*.. If System 1/2 leads to symbols, they'll likely be as abstraction to improve efficiency. There is little reason to expect that abstractions that a pure learning system creates on its own, much like Wittgenstein's Lion, will wind up aligning well with the ones we humans use.

up several significant challenges that need to be addressed by the research community: (1) the challenge of approximating the explanations–and constraining the interpretation of the human advice–in terms of the symbolic interface,(2) the challenge of assembling the symbolic interface itself–which in turn includes (2.1) getting the symbolic vocabulary and (2.2) grounding it in the representations that the AI system uses and (3) figuring out when and how to expand a pre-existing symbolic vocabulary to improve the accuracy of advice/explanation communication. In the remainder of the paper, we will discuss these challenges, and provide brief technical overviews of some existing research efforts–including some from our own research group–that can be viewed as initial realizations of such symbolic interface framework.

2 Research Challenges

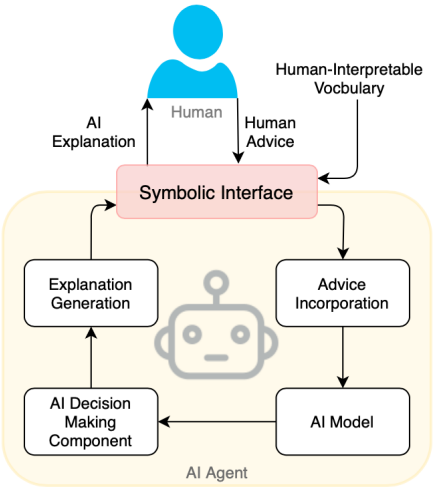

Figure 1: Overall architecture of an AI system exposing a symbolic interface to a human user, thereby allowing the AI agent to provide explanations to its decisions as well as accept guidance/preferences from the human in the form of advice.

Figure 1 presents an overview of an AI system capable of leveraging a symbolic interface of the type we advocate for. Note that in the architecture the agent's decision-making relies completely on its internal models, which may well be expressed, and operating on representations that are not directly accessible to the human; be they neural network based or based on some other internal symbolic abstraction. Note that the agent could interact with the human across multiple modalities (like annotated images, videos, etc, demonstrations, etc.). We are only arguing for the *inclusion* of a symbolic interface. We thus focus on symbolic interface expressed in human-understandable concepts. The AI system uses the symbolic interface to both communicate its explanations to the user and to receive instructions and advice from them. In this section, we will discuss in more detail how this symbolic interface could be used towards this end and also enumerate some of the open research challenges we need to address towards creating such systems.

2.1 Challenge: Interpreting Human Advice and Generating Explanation

Pending the details of how the symbolic interface is setup, let us first start by looking at how the interface is intended to be used by the AI system. In general, we expect the symbolic interface to allow the AI system to both capture and expose an approximate representation of parts of the system's model (or at least human's expectations of it) in terms the human can understand. Thus the symbolic interface can become the basis for providing explanations, wherein by exposing information about the underlying AI system in terms the user can understand, they can update their expectations about the AI system (thus performing model-reconciliation). By the same token, we can allow the human to provide updates for the AI system model in terms they understand for example, by specifying additional constraints, previously unspecified preferences, etc. This includes works like reward-machines (Icarte et al. 2018) or restraining bolts (De Giacomo et al. 2019), wherein the advice-giver is effectively updating the AI behavior by introducing these new trajectory preferences. Additionally, in the case of providing human advice, by the virtue of the symbolic interface being capable of capturing human expectations, it can be used as a tool to better contextualize, build on, and generalize human input. By leveraging the intuition that while coming up with the advice the human would have used a representation similar to the current interface, the AI system can potentially identify what objectives they may have had when providing the specific inputs.

2.2 Challenge: Collecting Initial Concept Set

The first point is to collect the propositional and relational concepts that will form the basis of the interactions (and potentially even action labels). These symbols are meant to form the conceptual representation of the task and or AI system's capabilities from the perspective of the human. Such symbols will be used to express any information the system may provide to the user and to analyze any input the user may provide. Note that in the case of user input, the input itself need not be expressed in symbolic terms, but the concepts provided allow the system to better utilize it. For example, as in the case of (Zha, Guan, and Kambhampati 2021), the agent may be provided a teleoperated demonstration by the user, but access to state factors that may be important to the user will allow the agent to better generalize the demonstration. These symbols may be obtained either directly or indirectly from the human. All the explicit knowledge representations – including knowledge graphs– will be in terms of symbols specified by a human. In a similar vein, extracting these symbols should also be feasible – for example the use of scene graph analysis (Krishna et al. 2017) – provided they are done from the human perspective. Also, note that our symbolic interfaces require more than just any symbolic representation. Works like (Konidaris, Kaelbling, and Lozano-Perez 2018), (Bonet and Geffner 2019) or (Ghorbani et al. 2019) that try to learn symbols from the perspective of the system need not result in useful symbolic interfaces, as those learned symbols may not make sense to humans. In gen-

eral, we will assume there exists a way to map the STST to the corresponding set of symbols. One way to accomplish this may be to learn specific classifiers that identify the presence or absence of individual concepts of interest in the given slice of STST. As mentioned earlier, we could also use methods like scene-graph analysis, that leverage the ability to identify common objects and their relationship to create a high-level symbolic representation of the relevant scene. For everyday scenarios, which do not require specialized vocabulary, such methods can be particularly powerful, as this allows us to use robust systems to generate symbolic representations without the additional overhead of collecting domain specific-data. We are thus effectively amortizing the cost of collecting the concepts over the life-time of all AI systems that use them to generate the symbolic interface. In the case where we are learning domain-specific vocabulary, the concept set itself could come from multiple sources, including the user of the system and system developers. Even in this case, we could try to amortize the concept collection cost by creating domain-specific concept databases, which could be used by multiple systems (and for multiple users). Potential concept lists could also be mined from documents related to the domains.

2.3 Challenge: Learning Concept Grounding

Now the next question would be how to learn the mappings between STST and symbolic concepts. The important point to recognize is that these groundings should try to approximate how the end-user would ground and understand the given concepts. In the case of off-the-shelf concept detection methods (like those that generate scene-graphs), the grounding is learned through large amounts of annotated data collected usually from crowd-sourced workers. In everyday scenarios, this is completely sufficient, as in general people tend to agree on the use of everyday concepts/words, etc (with possibly some cultural variations). On the other hand for more specialized domains, we may have to engage in a separate data-collection process to identify the grounding. In cases, where such mappings are captured through learned classifiers, this may require us to collect positive and negative examples from the concept specifier. If we are using a learning-based method, we would want to rely on existing few-shot learning methods to reduce the examples required per concept. If the agent is using learning methods that can compute its own representations of the state for coming up with decisions. Such simplified representations could then be used as the input to our vocabulary item classifiers.

Also note that any learned concept grounding is going to be noisy at best. This means that even if the concept classifiers or the scene-graph generator say certain concepts are present in a state or a slice of STST, it may not necessarily be true. Thus it may be extremely helpful to quantify the uncertainty over detected concepts. In the case of classifiers, one way to quantify such uncertainty may be to approximate the probabilistic accuracy of a given classifier (either estimated empirically over held-out data) or by using probabilistic classifiers. In cases, where we are generating entire scene graphs in addition to the uncertainty over the validity of the graph as a whole, it may be still helpful to quantify the certainty of specific objects and relationships in the scene graph. As one of the strengths of symbolic representation is modularity and composability, we hope to leverage it in the proposed system. So even when we can generate the complete scene graph we may choose to use or even expose only parts of the graph. Once estimated, such probabilistic beliefs could be explicitly taken into account by the algorithms that will be using these concepts. For example, if the concepts are going to be used to learn a symbolic model approximation, one could use a Bayesian approach and maintain multiple hypotheses over the possible models (with varying degrees of certainty). Additionally, if the system chooses to only expose the most likely model to the user, it can additionally also surface the certainty it has over that model.

2.4 Challenge: Allowing For Vocabulary Expansion

Another challenge that long term AI systems would have to deal with is the fact that the original concept list would, in most cases, be incomplete. So the algorithms that work with these concept lists will need to explicitly allow for the fact that they may only have a subset of the total vocabulary list that the human may have access to. This means that when it tries to reason about or interpret human input, or build a symbolic approximation of its own model, there may be concepts that are required for correctly performing such operations. In general, the problem of detecting incomplete vocabulary is easier for cases where the system is trying to build symbolic approximations (since it can approximate the accuracy of the learned representation), than in cases where it is using these concepts purely as a way to analyze the human input. As in the latter case, it is easy for the system to display confirmation bias and incorrectly adopt a hypothesis that is expressed solely in terms of the previously specified concepts. One way to avoid such biases may be to ensure that the system always maintains some uncertainty in any hypotheses it learns from the human input (Hadfield-Menell et al. 2017). Once vocabulary incompleteness is identified the next challenge is to work with the human to identify the missing vocabulary items. In this case, rather than trying to just acquire more concepts and symbols, the system could be more directed. Particularly in the case of explanations, the AI agent could rely on its underlying system model to identify what system states and model components may be relevant to the current explanatory queries. For example, in cases where the AI agent can generate visual representations of the underlying state, it can use low-level explanations like saliency maps to highlight parts of the state relevant to the current decisions. In the case of super-human AIs, there may be an additional challenge that the human vocabulary isn't sufficient to create a helpful explanation. In such cases, we may need to make use of strategies from intelligent tutoring systems (ITS) (Anderson, Boyle, and Reiser 1985) to enable the AI systems to teach concepts to humans.

3 Case Studies

3.1 Providing Explanation to Humans

In terms of works that have looked at the use of a symbolic interface for explanation generation, an instructive example would be (Sreedharan et al. 2020). The work builds on previous works like TCAV (Kim et al. 2018) that tries to identify the influence of propositional concepts on classifier decisions (thus building an abstract symbolic model that relates various concepts to system decisions). In (Sreedharan et al. 2020), given an explanatory query, usually, a contrastive one that asks why the current plan was selected over another that the user expected, the system tries to generate an explanation in terms the user understands. The decision-making system makes its decisions based on a model that is opaque to an end-user (say a learned model or a simulator). For a given explanatory query, the work tries to construct parts of a symbolic model (learned from samples generated from the opaque model), particularly missing preconditions and abstract cost function, expressed in human-understandable concepts. This model is then used to provide specific explanations. In regards to concept sets, the system assumes access to a set of user-specified propositional concepts, along with their corresponding classifiers. These classifiers are grounded based on positive and negative examples for each concept. The work captures the uncertainty regarding the grounding by using the classification accuracy of the system. Additionally, the system associates a level of uncertainty to each learned symbolic model-component, that not only captures any uncertainty related to grounding but also the fact that the system may have used too few samples to identify the correct model component. Finally, the system can detect that the original vocabulary set may be incomplete if the algorithm is unable to find an explanation for the user query.

3.2 Interpreting Human Advice

An illustration of the use of symbolic interfaces to more effectively interpret human advice is provided by EXPAND system (Guan et al. 2020), which tries to utilize human binary evaluative feedback and visual explanation to accelerate Human-in-the-Loop Deep Reinforcement Learning. The visual advice is given as saliency regions associated with the action taken. Such feedback could be expensive to collect, as well as unintuitive to specify, especially when the human has to provide such annotations for each query. To make the feedback collection process more efficient and effortless for the human expert, EXPAND leverages an object-oriented interface to convert the labels of relevant objects into corresponding saliency regions in image observations via off-the-shelf object detectors (which effectively "ground" human object symbols into the image STST). The human feedback is thus interpreted with the assumption that it refers to objects that are relevant from human's point of view. This despite the fact that the Deep RL part of EXPAND is operating over pixel space, and constructing its own internal representations. Object-oriented symbolic interfaces like these have also been used in other previous works to allow humans to provide informative and generalizable object-focused advices in an effortless way (Thomaz, Breazeal et al. 2006; Krening et al. 2016; Guan et al. 2020).

Another related recent project (Zha, Guan, and Kambhampati 2021) takes the same approach of using symbolic interface to better interpret the human advice. Here the aim is to reduce the ambiguity in human demonstrations of robotic tasks to improve the efficiency of reinforcement learning from demonstrations (RLfD). The system assumes that the (continuous) demonstration provided by the human is guided by their own interest in highlighting specific symbolic goals and way points. It learns to interpret the relative importance of these symbols and use that to disambiguate the demonstration (a process that can be viewed as the AI system trying to "explain" the demonstration to itself in terms of symbols that are viewed to be critical for the human demonstrator).

4 Concluding Remarks

In this paper, we argued for an ambitious research program focused on ensuring a symbolic interface to AI systems– independent of whether their internal operations themselves are done in human-interpretable symbolic means. In the end, the need to establish a human interpretable symbolic interface is but a natural consequence of the fact that even in the simplest case, human agent-interaction consists at the very least of three models (Kambhampati 2020). The human model $\mathcal{M}^H$ with which captures human capabilities, preferences, and objectives; the agent model $\mathcal{M}^R$ which drives the agent decision-making, and the pivotal bridge model $\mathcal{M}^R_h$ that captures human expectations of the agent. It is using $\mathcal{M}^R_h$ that the human decides what instructions and advice to give to the agent so they generate the desired outcomes, and it is by updating $\mathcal{M}^R_h$ that the robot could try to help the human make better sense of its behavior. It should be clear to the reader that in the most general case, our proposed symbolic interface acts as a way to facilitate the communication and manipulation of $\mathcal{M}^R_h$ in terms the human would have conceived it. When we use this symbolic interfaces to interpret the human to advise, we are but using the intuition that they were relying on $\mathcal{M}^R_h$ to come up with them in the first place. When we allow the user to manipulate the symbolic representation as part of advice, we are but allowing for the fact that the human expects the agent's model to be updated in response to their input. Finally, when we surface symbolic representations of the agent's true model, we are but performing model reconciliation with the added insight that the human mental model need not be represented in the same terms as the robot model. If we are to avoid the use of such a symbolic interfaces, we are doing so at the cost of our system's ability to be truly human-aware. Without such a interface expressed in terms human understands, we are effectively forcing the human to enter the world of the agent. Here it becomes the human's responsibility to make sure that their advice can be made sense in terms the agents can make sense of, and to take raw space time signal tubes emitted by the agent or corresponding to agent behavior and try to patch their existing expectations about the agent expressed in symbolic terms.

References

Anderson, J. R.; Boyle, C. F.; and Reiser, B. J. 1985. Intelligent tutoring systems. *Science* 228(4698): 456–462.

Bonet, B.; and Geffner, H. 2019. Learning first-order symbolic representations for planning from the structure of the state space. In *ECAI*.

De Giacomo, G.; Iocchi, L.; Favorito, M.; and Patrizi, F. 2019. Foundations for restraining bolts: Reinforcement learning with LTLf/LDLf restraining specifications. In *ICAPS*.

De Raedt, L.; Manhaeve, R.; Dumancic, S.; Demeester, T.; and Kimmig, A. 2019. Neuro-symbolic= neural+ logical+ probabilistic. In *NeSy'19@ IJCAI, the 14th International Workshop on Neural-Symbolic Learning and Reasoning*.

Garcez, A. d.; Gori, M.; Lamb, L. C.; Serafini, L.; Spranger, M.; and Tran, S. N. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088* .

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards automatic concept-based explanations. In *NeurIPS*, 9273–9282.

Greydanus, S.; Koul, A.; Dodge, J.; and Fern, A. 2018. Visualizing and Understanding Atari Agents. In *ICML*.

Guan, L.; Verma, M.; Guo, S.; Zhang, R.; and Kambhampati, S. 2020. Explanation augmented feedback in human-in-the-loop reinforcement learning. *arXiv preprint arXiv:2006.14804* .

Hadfield-Menell, D.; Dragan, A.; Abbeel, P.; and Russell, S. 2017. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*.

Icarte, R. T.; Klassen, T.; Valenzano, R.; and McIlraith, S. 2018. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*, 2107–2116. PMLR.

Kambhampati, S. 2020. Challenges of Human-Aware AI Systems AAAI Presidential Address. *AI Mag.* 41(3): 3–17. doi:10.1609/aimag.v41i3.5257. URL https://doi.org/10.1609/aimag.v41i3.5257.

Kambhampati, S. 2021. Polanyi's revenge and AI's new romance with tacit knowledge. *Commun. ACM* 64(2): 31–32.

Kim, B.; M., W.; Gilmer, J.; C., C.; J., W.; ; Viegas, F.; and Sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV) . *ICML* .

Konidaris, G.; Kaelbling, L. P.; and Lozano-Perez, T. 2018. From skills to symbols: Learning symbolic representations for abstract high-level planning. *Journal of Artificial Intelligence Research* 61: 215–289.

Krening, S.; Harrison, B.; Feigh, K. M.; Isbell, C. L.; Riedl, M.; and Thomaz, A. 2016. Learning from explanations using sentiment and advice in RL. *IEEE Transactions on Cognitive and Developmental Systems* 9(1): 44–55.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123(1): 32–73.

Sreedharan, S.; Soni, U.; Verma, M.; Srivastava, S.; and Kambhampati, S. 2020. Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations. *ICML-HILL Workshop* .

Thomaz, A. L.; Breazeal, C.; et al. 2006. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*, volume 6, 1000–1005. Boston, MA.

Zha, Y.; Guan, L.; and Kambhampati, S. 2021. Learning from Ambiguous Demonstrations with Self-Explanation Guided Reinforcement Learning. In *Submitted to ICRA 2022 and will be on arXiv e-prints soon*. Under review.

Zhang, R.; Walshe, C.; Liu, Z.; Guan, L.; Muller, K.; Whritner, J.; Zhang, L.; Hayhoe, M.; and Ballard, D. 2020. Atarihead: Atari human eye-tracking and demonstration dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6811–6820.